Cambridge Assessment

Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests

Conference Paper

Tom Bramley & Tom Benton

Presented at the 18th annual conference of the AEA – Europe conference
Prague, Czech Republic
8- 11 November 2017

Author contact details:

Tom Bramley & Tom Benton
Assessment Research and Development,
Research Division
Cambridge Assessment
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
UK

Bramley.T@cambridgeassessment.org.uk
Benton.T@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk

As a department of Cambridge University, Cambridge Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

How to cite this publication:

Bramley, T. & Benton, T. (2017). *Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests.* Paper presented at the 18th annual AEA-Europe conference, Prague, Czech Republic, 8-11 November 2017.

Abstract

If standard setting is conceived as a process whereby an abstraction (the performance standard) is made concrete as a cut-score on the raw score scale of a real test, then carrying out a standard-setting exercise on two tests is conceptually closely related to IRT true-score equating where score points on two tests corresponding to the same latent trait location are deemed equivalent. Noting that judge estimates of item difficulty typically correlate about the same with actual difficulty as empirical difficulty estimates based on very small samples (N<10), we compared, by simulation, a small-sample non-equivalent groups anchor test equating method with an Angoff-based method for determining equivalent cut-scores on two tests. At typical levels of correlation of judged and actual difficulty (r=0.6), small-sample equating with N=90 was more accurate than Angoff-based standard-setting. However, using a weaker anchor test or clustered sampling made the equating method similar to or worse than the Angoff-based method (depending on the cut-score location). We discuss implications for testing scenarios where these two approaches are likely to be feasible options.

Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests

Introduction

The educational measurement literature makes a clear distinction between the activities of standard setting on the one hand, and test equating or linking on the other. For example, these topics occupy different chapters in the standard reference work Educational Measurement (Brennan, 2006). Test equating is usually defined in a fairly narrow, technical way such as: "Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably." (Kolen & Brennan, 2004, p2). Standard setting, on the other hand is usually defined more broadly such as "…the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states of performance" (Cizek, 1993, p100). The main issues in test equating tend to be around the definition of the 'correct' equating transformation, and the data collection designs and statistical methods necessary to estimate it. In standard setting, however, the procedures are "… seldom, if ever, impartial psychometric activities, conducted in isolation. Social, political and economic forces impinge on the standard-setting process…" (Cizek & Earnest, 2015, p213). In particular, standard setting processes involve human values and judgments, and differences in these are to be expected.

However, the two different processes are clearly conceptually very closely related. This is brought out nicely by the graphic in Cizek & Earnest (ibid., p213) representing a 'performance standard' as a point (x) on a hypothetical performance continuum, and describing the standard setting process as being one of translating point x to a cut-score (point y) on the percentage correct scale of the test on which the standard is to be set. Thus the performance standard is an abstraction that could be realised (made concrete) by carrying out a standard setting exercise on any number of hypothetical tests. Doing this in practice on several actual tests, however, would doubtless lead to different cut-scores on the different tests, partly because tests constructed to the same specifications inevitably differ somewhat in difficulty. But it is this fact that creates the need for test equating processes! Although some test equating methods (those generally referred to as 'observed score equating methods') equate test forms without any reference to a hypothetical continuum, other methods, especially the IRT true score equating method, do make specific reference to such a continuum. In IRT true-score equating, the expected score on test X for an examinee with ability θ is deemed equivalent to the expected score on test Y for an examinee with ability θ, where θ is the examinee's 'ability' on the latent trait (abstract continuum) that is said to underlie performance on both tests.

If we are prepared to conceive of the abstract continuum on which the performance standard is located and the latent trait of the IRT model as one and the same, then we can see that carrying out separate standard setting exercises on tests X and Y is in theory no different from attempting to equate them (at the point on the latent trait corresponding to the cut-score) by a true-score IRT approach. Of course, the *results* of applying such dramatically different approaches to the same problem could be expected to differ.

Although it would seem most logically justifiable to carry out a standard setting exercise just once (to establish one definitive example of a realisation of the abstract performance standard on a concrete test) and then to use statistical equating to link all subsequent (or other) forms to that, in practice it may well be that a standard-setting method is used (perhaps alongside other methods) to inform or set the cut-score on subsequent forms. Thus the standard-setting method is used in practice as a test equating (standard-maintaining) method. There are several scenarios where this might arise, for example: i) if the test is very high-stakes (e.g. a licence-to-practise test) where procedures require 'stakeholder' involvement in setting the cut-score on each test form; ii) if sample sizes are so low on each test form that statistical equating methods are not trusted; iii) if contextual factors (such as cost, need for test security, local culture and expectations) prevent some of the necessities for equating methods such as pre-testing, administration of an anchor test, or embedding of field-test items into live tests; iv) if there is a need to determine a cut-score before any 'live' performance data has been collected. Interestingly, in a recent book Opposs & Gorgen (p57), deliberately chose to ignore the distinction between standard setting and standard maintaining because it was not helpful in considering the wide range of practices in different countries for arriving at cut-scores on high stakes tests.

The conceptual similarity between equating and standard setting raises questions of the relative accuracy of the two methods. Our starting assumption was that in an ideal world a large-sample equating exercise would be the preferred way to map a cut-score from one test to another parallel one. However, since the standard error of equating in a test equating exercise depends upon the sample size, continually reducing the sample size presumably will reach a point at which the equating error becomes greater than that of carrying out two separate standard setting exercises. The equating error from the latter will depend on the details of the method used but for all methods that rely on the judgment of item difficulty by experts a fundamental issue is the extent to which those judgments correspond to the actual empirical difficulty. One of the motivations for this research was the observation that the correlation of facility values[1] based on very small samples (N<10) with the facility values based on the full dataset of item scores are often of an order of magnitude similar to that reported for the correlation of judged difficulty with empirical difficulty. An example is given in Figure 1, which shows the distribution of correlations of facility values from 1,000 random samples of size N=1 to 10 with the correct facility value using the full sample (N=439) using publicly available data from a PISA mathematics test booklet.[2] The median correlation rises from around 0.5 to just under 0.9 as the sample size increases from 1 to 10, and we have found similar results from a variety of other data sets. Indeed, using the well-known formula relating standard errors of estimates of percentages to sample sizes, it is fairly easy to verify the expected value of these correlations mathematically for any set of facilities for one mark items and a given sample size. However, we felt it was more compelling to illustrate the correlations using a real data set. The crucial point is that the range of correlations is similar to those reported for the correlation between empirical difficulty and judgements of difficulty in the Angoff standard setting method. For example Brandon (2004) reports an average correlation of 0.61 (with an SD of 0.16) across 29 Angoff studies described in 7 research articles, and a more recent study by Tannenbaum & Kannan (2015) reported values around

[1] Facility value= mean item score / maximum possible item score

[2] The sixth test booklet from the assessment in 2012 for students in the USA. Available from https://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm.

0.5 to 0.8. These correlations are of the average estimated item difficulty across a number of judges with the empirical difficulty. The size of this correlation depends on the number of judges and the range of empirical item difficulty, as discussed in Thorndike (1982).

This similarity raises the question that prompted this research – given that estimates of item difficulty with similar correlations to the correct empirical value can be obtained from very small samples of empirical data, might it be the case that small-sample equating techniques are more effective than separate standard-setting exercises at mapping a cut-score from one test to another? In this paper we explore this question by simulation, but using data from a real assessment. The particular standard setting method and small sample equating method were chosen to reflect scenarios that we considered likely to occur in practice and hence to be relevant to a wide range of practitioners. They are described in more detail below.

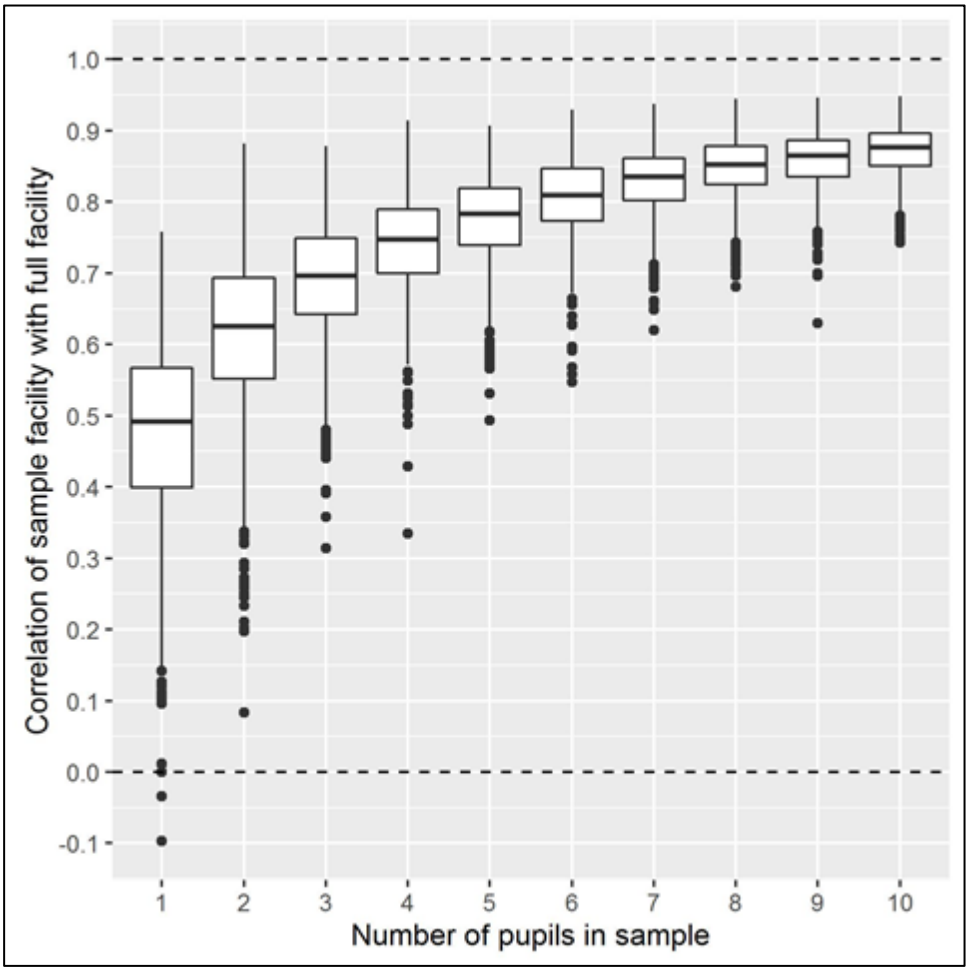

Figure 1. Distribution of correlations of facilities from small samples with overall facilities for booklet 6 mathematics items in PISA 2012 for US pupils.

Standard setting method 'Mean estimation'

The mean estimation method is a variant of the more well known Angoff method (e.g. Angoff, 1970; Loomis & Bourque, 2001). It is applicable to tests containing polytomous as well as dichotomous items. If the test consists solely of dichotomous items it is the same as the Angoff method. Experts estimate the difficulty of each of the items in a test in terms of the mean score likely to be obtained on each item by a group of minimally competent examinees

(MCEs). If the test is pass-fail then the MCEs are those who are just competent enough to pass. If the test is graded into more than two categories there are different groups of MCEs for each cut-score. The cut-score is derived by summing the estimated means and then averaging across judges, rounding the result to an integer if necessary (or averaging and then summing – it makes no difference). Hence tests comprising items that are perceived to be more difficult will have cut-scores set at lower values. There are various procedural variations possible, such as allowing rounds of judgements where the experts discuss and compare their estimates before making a further set of estimates; or allowing the experts to see empirical evidence (perhaps from pre-testing) about the difficulty of the items; or allowing the experts to see the effect of their judgements on the pass rate (or grade distribution) of the test.

Note that previous research (e.g. Impara and Plake, 1998) has suggested that although estimating the mean scores of MCEs can be difficult for experts in an absolute sense, they are more adept at discerning the correct rank order of the difficulty of items. Hence judgments from experts can potentially be transformed onto the correct scale before being used to inform standard setting (Thorndike, 1982; Humphry, Heldsinger, and Andrich, 2014). Since judgments can be transformed to the correct scale, the correlation between estimated difficulties and actual difficulties (often measured by item facilities) provides a reasonable idea of the value of the information from such methods, as discussed above. In our simulation (described in more detail later) we wanted to vary this level of correlation and assess the effect on the outcome.

Small-sample equating method: chained linear equating

There are a variety of equating methods appropriate for use with small samples (for example, see Livingston and Kim, 2009, or Kim, von Davier and Haberman, 2008). The experimental design is relevant to the choice of method: we wanted a method suitable for the 'non-equivalent groups anchor test' (NEAT) design. This is because for equating test forms which are produced once a year (for example) it is often not possible to get one group of examinees to take both forms, or to obtain randomly equivalent groups of examinees. Much more often it is possible to obtain two different groups and adjust statistically for differences in ability between them by means of an anchor test or other covariate. Just as the correlation between judged and empirical difficulty is relevant for the validity of standard setting methods, so the correlation of test scores with anchor test scores is relevant for the validity of equating methods using an anchor test. In our simulation we wanted to use an anchor with a higher and a lower correlation. We chose chained linear equating (e.g. Puhan, 2010) because it requires fewer parameters to be estimated than the theoretically preferable (with large samples) equipercentile equating, and because one of our anchor tests was a continuous covariate for which we could calculate mean and standard deviation (necessary for linear equating) but where there was no discrete score frequency distribution. Puhan (2010) also reports that, across a range of conditions, chained linear equating tends to perform well compared to other linear equating techniques for the NEAT design.

As well as exploring the effect of the correlation of the anchor test, we were also interested in exploring the effect of clustering. In practice it might only be logistically feasible to obtain examinees from a single class in a small number of schools for an equating exercise, so it

was of interest to see how a small clustered sample differed from a genuinely random sample of the same size.

In brief, the equating scenario consisted of a test X (where we assumed the cut-scores were known) and a test Y where we needed to set equivalent cut-scores. We simulated mean-estimation judgments at two levels of correlation (0.6 and 0.9) between estimated and empirical values, and derived the cut-score on test Y by adding up the simulated means for the items on test Y. We compared this with a chained linear equating method in three conditions:

1) Random samples of 30 examinees from three schools in group A took test X and from three schools in group B took test Y, and all 180 examinees took an anchor test V;
2) Simple random samples of 90 examinees in group A took test X and in group B took test Y, and all 180 examinees took an anchor test V;
3) The same random samples as in 2) were used to equate with a weaker anchor test W, where 'weaker' implies that scores on W were less highly correlated with scores on X and Y than scores on V were.

Condition 1 corresponds to a scenario where one class of 30 pupils in three schools can be induced or made to take an anchor test each year. The difference between condition 2 and 1 reflects the effect of clustering on equating error. The difference between condition 3 and 2 reflects the effect of the power of the anchor test to adjust for differences in examinee ability. In all cases we considered two cut-scores, one at the lower end of the raw score scale and one at the higher end.

Method

Data

The dataset forming the basis of all the analyses reported here was artificially constructed from a large real dataset containing the responses of 15,731 examinees to a test with a maximum possible raw score of 200. The questions were made up of sub-questions (henceforth items), and the items ranged in tariff (maximum score) from 1 (i.e. dichotomous) to 5 (i.e. polytomous with 6 score categories). The facility values of all items were calculated and two tests X and Y, each with a maximum possible raw score of 60, were constructed by selecting two sets of items comprising fifteen 2-tariff items and ten 3-tariff items by systematically alternating selection from the items ordered by facility value. An anchor test V was constructed from 20 dichotomous items (which was all the dichotomous items and hence no selection method was required).

The examinees came from 323 centres in total (the vast majority of which were schools and are for simplicity referred to henceforth as schools), each contributing between 1 and 238 examinees (mean 48.7, median 33). Each school had a 5-digit identification number, which was known to be non-randomly assigned. Two non-equivalent groups of examinees of roughly the same size were created by assigning those in schools with ID numbers below a certain value to Group A and the rest to Group B.

Table 1 shows that test Y was slightly easier than test X (mean score higher) but the lower SD of scores on test Y shows that the difference in difficulty was not uniform across the score range. It is also clear that Group A was of higher ability than Group B (mean score higher on all tests).

Table 1: Descriptive statistics for scores on tests X, Y and V.

	All (N=15,731)		*Group A (N=7,752)*		*Group B (N=7,979)*	
Test	Mean	SD	Mean	SD	Mean	SD
X (max 60)	31.76	13.29	33.00	13.39	30.55	13.08
Y (max 60)	32.36	12.16	33.46	12.37	31.29	11.86
V (max 20)	10.01	3.44	10.30	3.50	9.72	3.34

Definition of the correct equating function

As described in the introduction, for theoretical reasons we defined the 'correct' equating function to be the one arising from IRT true score equating on the complete dataset (i.e. X,Y and V items calibrated concurrently for both groups). The item calibrations were obtained by fitting the Rasch Partial Credit model (Masters, 1982) using the eRm package in R (Mair & Hatzinger, 2007). In this paper we consider two different cut-scores on test X: 15 out of 60, and 45 out of 60. The 'definitive' equated cut-scores on test Y arising from the IRT true score equating were 17.20 and 44.21.

Simulating estimated difficulty judgments

From the IRT analysis we know the ability θ corresponding to cut-scores of 15 or 45 on test X and hence[3] the expected scores of examinees with this ability on each item on test X and test Y. The sums of these expected scores over the items on test X are 15.00 and 45.00, and on test Y they are 17.20 and 44.21. These item-level expected scores are taken to be the empirical values that the judges are attempting to estimate and we refer to them later in this paper as EMs (Expected Means)[4]. The index of agreement most commonly reported in the standard-setting literature is the correlation of estimated item means (or p-values for tests containing only dichotomous items) with empirical means (which are monotonically related to the empirical mean score of MCEs). We simulated two levels of correlation: 0.6 (a value representative of published Angoff studies), and 0.9 (a much higher value than usually found, in order to represent a very optimistic view of what might be achievable in ideal conditions).

The simulation was carried out as follows:

1. The model-based expected item means for examinees at the cut-score (EM) for all items on test X and Y were converted to p-values (EP) (i.e. between 0 and 1) by dividing the mean by the maximum possible score for the item.
2. These p-values were converted to a logit scale of 'easiness' (EL) by the applying the transformation EL=log[EP/(1-EP)]. The mean and standard deviation (SD) on this scale are denoted *m* and *s* below.

[3] Via the equation for the Rasch Partial Credit model

[4] Of course, the actual empirical values would differ to the extent that the empirical IRFs deviated from the model-based item response functions (IRFs). Inspection of these IRFs showed close agreement for the vast majority of items. Graphs of the IRFs are available from the authors but not included to save space.

3. Estimated judgments on the logit scale (JL) with a specified (average) correlation *r* were created by combining the logit facilities from step 2 with a normally distributed random error component *e* ~ N(0,*s*) according to the formula:

$$JL = r \times EL + \sqrt{(1 - r^2)} \times e + (1 - r) \times m \qquad (1)$$

Error was added to the EL scale rather than to the EM scale to i) allow for the fact that the EM scale is bounded between 0 and the maximum score and hence error is unlikely to be normally distributed around the empirical value when it is close to these bounds; and ii) to allow for the fact that there is more potential for error when judging an item with a larger maximum possible score. Note that the simulation created (on average) a JL scale with the same mean and standard deviation as the EL scale, a point we return to in the discussion.
4. These estimated judgments were converted back to judged p-values (JP) and judged item means (JM) by applying the reverse transformations to those described in steps 1 and 2.
5. This process was repeated 1,000 times for each of two different values of the correlation *r* (0.9 and 0.6) and for two different test X cut-scores (15 and 45).
6. For each replicate, cut-scores on test Y were derived by summing the JM values for the test Y items.
7. The distributions of equated cut-scores were compared with the definitive (correct) cut-score. Specifically, bias B was defined as the mean difference (across replicates) between the equated score for each replicate and the correct cut-score; error variance E was defined as the variance of the equated cut-scores; and the root mean squared error RMSE was calculated as sqrt(B^2+E).

Creating replicates for small-sample equating

For condition 1, all schools with 30 or more examinees were selected and then a two-stage sampling process first selected at random three schools from each group, and then a random sample of 30 examinees from each school. This process was replicated 1,000 times. An equated cut-score on test Y for each of the test X cut-scores (15 and 45) was derived by chained linear equating in each replicate.

For around 70% of examinees in the main dataset we had information about scores on other assessments which we were able to combine into a single measure of general ability (Benton, 2017). For the purpose of this research we treated this measure as if it were another anchor test W, less strongly correlated with scores on test X and Y, as shown in Table 2.

Table 2: Pearson correlations of scores on tests X, Y, V and W.

	Complete data (Condition 1)			*Examinees with score on W (Conditions 2 & 3)*		
	All	Group A	Group B	All	Group A	Group B
Number of examinees	15,731	7,752	7,979	11,078	6,311	4,767
X with Y	0.900	0.903	0.894	0.895	0.900	0.885
X with V	0.812	0.818	0.803	0.804	0.813	0.789
Y with V	0.820	0.826	0.812	0.813	0.819	0.800
X with W	-	-	-	0.640	0.633	0.643
Y with W	-	-	-	0.624	0.618	0.629

We selected 1,000 simple random samples (with replacement) of 90 examinees from Group A and 90 from Group B out of the 11,708 examinees with a score for both V and W. For

condition 2 an equated cut-score on test Y for each of the test X cut-scores (15 and 45) was derived by chained linear equating with anchor test V in each of the 1000 samples. Condition 3 used the same samples but equated via the anchor test W instead of V. In all conditions the chained linear equating transformation was calculated as:

$$cl_Y(C_X) = \frac{s_{YB}}{s_{VB}} \times \frac{s_{VA}}{s_{XA}}(C_X - m_{XA}) + m_{YB} + \frac{s_{YB}}{s_{VB}}(m_{VA} - m_{VB}) \qquad (2)$$

where:

$cl_Y(C_X)$ is the chained linear equated score on test Y for a cut score C_X on test X,
m_{VA} is the mean score on anchor test V from the schools in Group A,
s_{VA} is the SD of scores on anchor test V from the schools in Group A,
and other mean and SD terms are likewise defined (see Albano, 2016, equation 70).
The distribution of equated scores across the 1,000 replicates was then compared with the definitive cut-score in the same way as for the simulated judgments.

Results

The results for the simulated judgments are shown in Table 3. In all cases the bias made a negligible contribution to the overall RMSE. The more realistic value for the correlation (0.6) had RMSE values nearly twice as high as that for the optimistic value (0.9) at both cut-scores. The % distributions in Table 3 refer to equated cut-scores on test Y rounded to the nearest integer. This is on the assumption that in practice if an integer cut-score were required to be set on test Y, the correct values would be 17 and 44. This causes a slight asymmetry because an equated score of 44.6 (say) would be rounded to 45 and be 1 too high, whereas a less accurate equated score of 43.6 would be rounded to the correct value of 44. For simulated correlations of 0.9, the equated cut-score was within ±1 of the correct score around 75% of the time (cut-score of 15) or 80% of the time (cut-score of 45), but for simulated correlations of 0.6 only around 50% were in this range, and around 25% were 3 or more score points away.

Table 3: Equated scores based on simulated judgments (replications=1,000).

Test X cut-score	*15*	*15*	*45*	*45*
Simulated correlation	0.6	0.9	0.6	0.9
Mean (SD) correlation - logits	0.61 (0.12)	0.91 (0.03)	0.59 (0.13)	0.90 (0.03)
Mean (SD) correlation - means	0.54 (0.15)	0.87 (0.05)	0.72 (0.10)	0.92 (0.03)
Correct Y cut-score	17.20	17.20	44.21	44.21
Test Y mean equated cut-score	17.08	17.15	44.38	44.33
Test Y SD equated cut-score	2.30	1.25	2.06	1.12
Bias	-0.12	-0.05	0.18	0.12
RMSE	2.31	1.25	2.07	1.13
%<= -3	12.1	0.9	8.8	0.9
% -2	13.3	8.1	9.3	4.3
% -1	16.4	21.9	14.3	17.6
% 0	17.6	29.6	18.4	32.0
% +1	15.0	25.6	19.4	31.0
% +2	10.3	11.3	13.8	12.0
% >= +3	15.3	2.6	16.0	2.2

Table 4 shows that the overall precision of small sample equating, as measured by the RMSE, was best at both cut-scores in Condition 2 (simple random sample of 90 examinees from each test and a strong anchor). At both cut-scores the Condition 2 RMSE was roughly half-way between those from simulated judgments of 0.6 and 0.9. At both cut-scores the Condition 1 RMSE was about 0.7 score points higher (less precise) than that in Condition 2, showing the detrimental effect of clustering of examinees within schools on equating error. The Condition 1 RMSEs were slightly higher than those from simulated judgments with a correlation of 0.6. At a cut-score of 45 the Condition 3 RMSE was about 0.5 score points higher than Condition 2, but at a cut-score of 14 it was 1.4 score points higher, showing the detrimental effect of using a weaker (less strongly correlated) anchor test. In the best case for small sample equating (Condition 2) the cut-scores were within 1 score point of the correct value around 60% of the time for a cut-score of 15 and around 70% of the time for a cut-score of 45. Bias made a very small contribution to the RMSE at a cut-score of 15 and a negligible contribution at a cut-score of 45. The fact that sampling error was the main contributor to RMSE in all methods and conditions suggests that comparisons are not critically dependent on how the 'true' equating function is defined, because this would only affect the bias and not the sampling error.

Table 4: Equated scores based on small sample equating (replications=1,000).

Condition	*1*	*2*	*3*	*1*	*2*	*3*
Test X cut-score	15	15	15	45	45	45
Correct Y cut-score	17.20	17.20	17.20	44.21	44.21	44.21
Test Y mean cut-score	16.39	16.56	16.31	44.25	44.36	44.22
Test Y SD cut-score	2.41	1.70	3.08	2.18	1.45	1.94
Bias	-0.81	-0.64	-0.89	0.04	0.15	0.01
RMSE	2.54	1.82	3.21	2.18	1.45	1.94
%<= -3	19.1	12.0	26.2	10.5	1.9	6.6
% -2	10.4	13.6	12.3	9.9	7.9	13.8
% -1	18.8	22.1	10.8	16.9	17.5	16.7
% 0	19.7	23.2	12.7	19.0	27.1	20.0
% +1	14.1	16.0	13.3	16.2	24.6	19.1
% +2	10.3	9.7	10.9	11.4	13.4	12.2
% >= +3	7.6	3.4	13.8	16.1	7.6	11.6

Discussion

This study has compared, by simulation, the level of accuracy and precision that might be obtained from a standard-setting method (mean estimation) if applied as a test equating method to that which might be expected from a small-sample test equating method (chained linear equating). As expected, for the standard-setting method more accurate equating arose from a higher level of correlation between simulated expert judgments of item difficulty and empirical difficulty. For small sample equating with 90 examinees per test, more accurate equating arose from: i) using simple random sampling compared to cluster sampling at a given sample size; and ii) using a stronger rather than a weaker anchor. The actual values of RMSE depended on the cut-score, being generally larger for the cut-score where the correct equated cut-score on test Y was further from the cut-score on test X. The simulations based on the more realistic value for the correlation between judged and empirical difficulty (0.6) produced a higher value for the RMSE than the small-sample equating with random sampling and a strong anchor, and a similar RMSE to small-sample equating with cluster sampling. The comparison with random sampling and a weak anchor depended on the cut-score, with equating having the lower RMSE at a cut-score of 45 but standard-setting having the lower RMSE at a cut-score of 15. Simulations of standard-setting based on the optimistic correlation of 0.9 had the lowest RMSEs of all.

Given that, as seen earlier in this paper, even very small samples of examinees can give a more accurate picture of the relative difficulty of items than estimates from experts, we may be surprised that the small sample approach trialled here did not perform even better. There are a number of reasons for this. One reason is that the equating approach adopted in the simulation study required calibration of examinee abilities across two groups using an anchor test. Small sample equating with a single group design would be significantly more accurate. Even within the NEAT design, it may be that other approaches such as Tucker linear equating or Rasch true score equating may provide a more stable estimates of equivalent scores than chained linear equating. Finally, it is important to note that our simulations assumed that judged and empirical values for the mean scores of MCEs would differ only in their rank order, and that the mean and SD would (apart from sampling error) be the same. In fact evidence both old (Lorge & Kruglov, 1953) and new (Humphry, Heldsinger & Andrich, 2014) suggests that expert judges tend to think that easy items are harder than they are, and that hard items are easier than they are. That is, the implied scale unit of estimated difficulty tends to be larger (i.e. less discriminating) than the scale unit of empirical difficulty: the judges' estimates are less spread out than the empirical values. Humphry et al. (ibid). suggested applying a linear transformation (on the logit scale) to align the scale units, on the assumption that judges are unbiased when estimating passing proportions/probabilities of 50%. Although this assumption seems reasonably plausible it nevertheless needs empirical support. In any event, we were not confident that we could choose realistic values for scale shrinkage effects to include in our simulation because they may depend on a number of contextual factors. This is an area for further research.

In our simulations, sampling error was the dominant contributor to RMSE, which suggests that attempting to reduce sampling error at the risk of increasing bias may also be worth considering. One way of achieving this would be to apply the 'synthetic linking' approach of Kim, von Davier & Haberman (2008) where the final equated cut-score on test Y is a

weighted average of the test X cut-score and the cut-score derived from the equating. This approach is clearly most suitable when there is some reason to believe that the two tests should have similar cut-scores – perhaps if they have been constructed to the same detailed specification.

The main issue is whether the aggregate of judges' estimates of item difficulty provides useful information about relative test difficulty. The degree of correlation between judged and empirical item difficulty is clearly an important factor in reducing the RMSE of Angoff-related standard-setting methods. Using a small-sample equating method may be preferable to using a standard-setting method if typical levels of correlation are to be expected, and indeed this was the conclusion of Dwyer (2016), although it should be noted that the (actual, not simulated) correlations of the judge estimates in his study were in the range 0.39 to 0.49 – lower than observed in many other studies. However, if it is possible to increase the correlation beyond 0.6 by increasing the number of judges in a judging panel and/or training them to make the mean-estimation judgments then substantial improvements in the accuracy of the standard-setting method could be obtained – in the scenario considered here a correlation of 0.9 was more accurate than the best small sample equating (a random sample of 90 examinees and a strong anchor).

In conclusion, it can be observed that in some contexts standard setting methods are used to achieve the same goal as test equating methods, namely determining cut-scores on test forms that relate to the same performance standard. IRT true-score equating provides a conceptual link between the two if it is reasonable to conceive of the IRT latent trait as being the same as the abstract continuum containing the performance standard. The simulations reported here have suggested that the precision (RMSE) of Angoff-based standard setting methods could in some circumstances be similar to what might be expected from test equating with a NEAT design using small samples (N~100) of examinees. Of course, these findings all derive from simulations based on just one dataset, so we are not in a position to make general recommendations about what to do in particular applied contexts. We made choices about how to define the 'true' equating function, and which particular standard-setting method and small-sample equating method to use, all of which could be varied. The effect of using polytomous items rather than dichotomous anchor items could be explored, as could the effect of varying test length. Furthermore, our method of artificially constructing tests X and Y ensured that they would be reasonably similar in difficulty. However, these findings point to a way in which practitioners could set up experiments or simulations that more closely match their own particular contexts in order to discover whether using a standard-setting method based on expert judgment might be more accurate than using a small-sample test equating method (or vice versa); or whether focusing effort on constructing parallel (equally difficult) tests would be a better use of available resource.

References

Albano, A. D. (2016). equate: an R package for observed-score linking and equating. *Journal of Statistical Software, 74 (8)*.

Benton, T. (2017). *Pooling the totality of our data resources to maintain standards in the face of changing cohorts.* Paper presented at the 18th annual AEA-Europe conference, Prague, Czech Republic, 9-11 November 2017.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*(1), 59-88.

Brennan, R. L. (Ed.) (2006). *Educational Measurement* (4th ed.). Washington, DC: American Council on Education / Praeger.

Cizek, G. J. (1993). Reconsidering Standards and Criteria. *Journal of Educational Measurement, 30*(2), 93-106.

Cizek, G. J., & Earnest, D. S. (2015). Setting performance standards on tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 212-237). New York: Routledge.

Dwyer, A. C. (2016). Maintaining equivalent cut scores for small sample test forms. *Journal of Educational Measurement, 53*(1), 3-22.

Humphry, S., Heldsinger, S., & Andrich, D. (2014). Requiring a consistent unit of scale between the responses of students and judges in standard setting. *Applied Measurement in Education, 27*(1), 1-18.

Kim, S., von Davier, A. A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement, 45*(4), 325-342.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* (2nd ed.). New York: Springer.

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement*, 35, 69–81.

Livingston, S.A., & Kim, S, (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement*, 46, 330-343.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives* (pp. 175-217). Mahwah, NJ: Lawrence Erlbaum Associates.

Lorge, I., & Kruglov, L. (1953). The improvement of estimates of test difficulty. *Educational and Psychological Measurement, 13*, 34-46.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1-20. Retrieved from http://www.jstatsoft.org/v20/i09/paper

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Opposs, D., & Gorgen, K. (2018). What is standard setting? In J.-A. Baird, T. Isaacs, D. Opposs, & L. Gray (Eds.), *Examination standards: how measures and meanings differ around the world* (pp. 54-76). London: UCL Institute of Education Press.

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement*, *47*(1), 54-75.

Tannenbaum, R. J., & Kannan, P. (2015). Consistency of Angoff-based standard-setting judgments: are item judgments and passing scores replicable across different panels of experts? *Educational Assessment, 20*(1), 66-78.